UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

RAW STORY MEDIA, INC., ALTERNET MEDIA, INC.,

Plaintiffs,

v.

OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC,

Defendants.

Civil Action No. 24-01514-CM

FIRST AMENDED COMPLAINT

JURY TRIAL DEMANDED

1. Plaintiffs Raw Story Media, Inc. and AlterNet Media, Inc., through their attorneys Loevy & Loevy, for their Complaint against the OpenAI Defendants, allege the following:

2. The Copyright Clause of the U.S. Constitution empowers Congress to protect works of human creativity. The resulting legal protections encourage people to devote effort and resources to all manner of creative enterprises by providing confidence that creators' works will be shielded from unauthorized encroachment.

3. In recognition that emerging technologies could be used to evade statutory protections, Congress passed the Digital Millennium Copyright Act in 1998. The DMCA prohibits the removal of author, title, copyright, and terms of use information from protected works where there is reason to know that it would induce, enable, facilitate, or conceal a copyright infringement. Unlike copyright infringement claims, which require copyright owners to incur significant and often prohibitive registration costs as a prerequisite to enforcing their copyrights, a DMCA claim does not require registration.

4. Generative artificial intelligence (AI) systems and large language models (LLMs) are trained using works created by humans. AI systems and LLMs ingest massive amounts of human creativity and use it to mimic how humans write and speak. These training sets have included hundreds of thousands, if not millions, of works of journalism.

5. Defendants are the companies primarily responsible for the creation and development of the highly lucrative ChatGPT AI products. According to the award-winning website Copyleaks, nearly 60% of the responses provided by Defendants' GPT-3.5 product in a study conducted by Copyleaks contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content.

6. When they populated their training sets with works of journalism, Defendants had a choice: they could train ChatGPT using works of journalism with the copyright management information protected by the DMCA intact, or they could strip it away. Defendants chose the latter, and in the process, trained ChatGPT not to acknowledge or respect copyright, not to notify ChatGPT users when the responses they received were protected by journalists' copyrights, and not to provide attribution when using the works of human journalists.

7. Plaintiffs Raw Story and AlterNet are news organizations, and bring this lawsuit seeking actual damages and Defendants' profits, or statutory damages of no less than $2500 per violation.

PARTIES

8. For over two decades, Raw Story has published award-winning investigative journalism, breaking news, and bold opinion columns. Raw Story publishes to more than ten million readers each month and has more than 1,000,000 daily readers. It is the largest independent

progressive political news website in America and was named the best news/political blog in America by *Editor & Publisher* in 2022 and 2023.

9. Among other important work, Raw Story has received *Editor & Publisher* (EPPY), Society of Professional Journalists, Fair Media Council, and ION awards for its reporting on white nationalism, the January 6 riots, South Dakota governor Kristi Noem's use of a state airplane for non-official purposes, and inappropriate Congressional stock trading. Raw Story reporters produce timely, illuminating work at great risk and cost to enrich understanding of critical issues and undermine threats to civil society.

10. As one example of the risks taken by Raw Story reporters in bringing the news to the public—risks never faced by AI bots—members of a neo-Nazi group showed up at the home of a Raw Story reporter who covers extremism and white supremacy in America. *See* Washington Post, *A reporter investigated neo-Nazis. Then they came to his house in masks.* (Feb. 20, 2024), https://www.washingtonpost.com/style/media/2024/02/20/raw-story-neo-nazi-journalist-house/.

11. Raw Story is a Massachusetts corporation with its headquarters in Miami Beach, Florida.

12. AlterNet is a three-time Webby award-winning publisher with a focus on civil rights, social justice, culture, health, and the environment. For 25 years, AlterNet's reporters have chased political news, and its opinion writers have probed the intersection of politics, science, and religion.

13. AlterNet is a Delaware corporation with its headquarters in Miami Beach, Florida.

14. Together, Plaintiffs have published more than 400,000 breaking news features, investigative news articles and opinion columns as a result of their considerable investments of time and resources.

15. Defendants are the inter-related organizations primarily responsible for the creation, training, marketing, and sale of ChatGPT AI products.

16. OpenAI Inc. is a Delaware nonprofit corporation with a principal place of business in San Francisco, CA.

17. OpenAI OpCo LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. OpenAI OpCo LLC is the sole member of OpenAI, LLC. Previously, OpenAI OpCo was known as OpenAI LP.

18. OpenAI GP, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It is the general partner of OpenAI OpCo and controls OpenAI OpCo.

19. OpenAI, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It owns some of the services or products operated by OpenAI.

20. OpenAI Global LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA.

21. OAI Corporation, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole member is OpenAI Holdings, LLC.

22. OpenAI Holdings, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole members are OpenAI, Inc. and Aestas Corporation.

JURISDICTION AND VENUE

23. The Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, et seq., as amended by the Digital Millennium Copyright Act.

24. Jurisdiction over Defendants is proper because they have purposefully availed themselves of New York to conduct their business. OpenAI maintains offices and employs staff in New York who, on information and belief, were engaged in training and/or marketing OpenAI's GenAI systems and LLMs, and thus in the removal of Plaintiffs' copyright management information as discussed in this Complaint and/or the sale of products to New York residents resulting from that removal. Defendants did not contest personal jurisdiction in their Motion to Dismiss.

25. Venue is proper under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District.

26. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the acts or omissions giving rise to Plaintiffs' claims occurred in this District. Specifically, OpenAI employs staff in New York who, on information and belief, were engaged in the activities alleged in this Complaint.

27. Defendants did not contest venue in their Motion to Dismiss.

PLAINTIFFS' COPYRIGHT-PROTECTED WORKS OF JOURNALISM

28. Plaintiffs' copyrighted works of journalism are published on Plaintiffs' websites, rawstory.com and alternet.org respectively, and are conveyed to the public with author, title, and copyright notice information.

29. Plaintiffs own copyrights to all the Raw Story articles listed in Exhibit 1 and the AlterNet articles listed in Exhibit 2.

30. Plaintiffs' copyright-protected works are the result of significant investment by Plaintiffs in the human and other resources necessary to report on the news.

DEFENDANTS' INCLUSION OF PLAINTIFFS' WORKS IN THEIR TRAINING SETS AND REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION

31. Defendants' generative AI products utilize a "large language model," or "LLM." The different versions of GPT are examples of LLMs. An LLM, including those that power ChatGPT, take text prompts as inputs and emit outputs to predict responses that are likely to follow a given the potentially billions of input examples used to train it.

32. LLMs arrive at their outputs as the result of their training on works written by humans, which are often protected by copyright. They collect these examples in training sets.

33. When assembling training sets, LLM creators, including Defendants, first identify the works they want to include. They then encode the work in computer memory as numbers called "parameters."

34. Defendants have not published the contents of the training sets used to train any version of ChatGPT, but have disclosed information about those training sets prior to GPT-4.[1] Beginning with GPT-4, Defendants have been fully secret about the training sets used to train that and later versions of ChatGPT. Plaintiffs' allegations about Defendants' training sets are therefore based upon an extensive review of publicly available information regarding earlier versions of ChatGPT and consultations with a data scientist employed by Plaintiffs' counsel to analyze that information and provide insights into the manner in which AI is developed and functions.

35. Plaintiffs' allegations about Defendants' training sets are also based upon information learned through discovery in this case.

36. Earlier versions of ChatGPT (prior to GPT-4) were trained using at least the following training sets: WebText, WebText2, and sets derived from Common Crawl.

[1] Plaintiffs collectively refer to all versions of ChatGPT as "ChatGPT" unless a specific version is specified.

37. WebText and WebText2 were created by Defendants. They are collections of all outbound links on the website Reddit that received at least three "karma."[2] On Reddit, a karma indicates that users have generally approved the link. The difference between the datasets is that WebText2 involved scraping links from Reddit over a longer period of time. Thus, WebText2 is an expanded version of WebText.

38. Defendants have published a list of the top 1,000 web domains present in the WebText training set and their frequency. According to that list, WebText included 33,598 distinct URLs from Raw Story's web domain and 23,183 distinct URLs from AlterNet's web domain.[3]

39. Defendants have a record of, and are aware, of each URL that was included in each of their training sets.

40. Joshua C. Peterson, currently an assistant professor in the Faculty of Computing and Data Sciences at Boston University, and two computational cognitive scientists with PhDs from U.C. Berkeley, created an approximation of the WebText dataset, called OpenWebText, by also scraping outbound links from Reddit that received at least three "karma," just like Defendants did in creating WebText.[4] They published the results online. A data scientist employed by Plaintiffs' counsel then analyzed those results. OpenWebText contains 40,403 distinct URLs from rawstory.com and 34,003 from alternet.org. A list of the Raw Story works contained in OpenWebText is attached as Exhibit 3. A list of the AlterNet works contained in OpenWebText is attached as Exhibit 4.

---

[2] Alec Radford et al, Language Models are Unsupervised Multitask Learners, 3, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[3] https://github.com/openai/gpt-2/blob/master/domains.txt.

[4] https://github.com/jcpeterson/openwebtext/blob/master/README.md.

41. Upon information and belief, there are different numbers of Plaintiffs' articles in WebText and OpenWebText at least in part because the scrapes occurred on different dates.

42. Defendants have explained that, in developing WebText, they used sets of algorithms called Dragnet and Newspaper to extract text from websites.[5] Upon information and belief, Defendants used these two extraction methods, rather than one method, to create redundancies in case one method experienced a bug or did not work properly in a given case. Applying two methods rather than one would lead to a training set that is more consistent in the kind of content it contains, which is desirable from a training perspective.

43. Dragnet's algorithms are designed to "separate the main article content" from other parts of the website, including "footers" and "copyright notices," and allow the extractor to make further copies only of the "main article content."[6]

44. Like Dragnet, the Newspaper algorithms are incapable of extracting copyright notices and footers. Further, a user of Newspaper has the choice to extract or not extract author and title information. On information and belief, Defendants chose not to extract author and title information because they desired consistency with the Dragnet extractions, and Dragnet is designed not to extract author and title information.

45. In applying the Dragnet and Newspaper algorithms while assembling the WebText dataset, Defendants removed Plaintiffs' author, title, and copyright notice information.

---

[5] Alec Radford et al., Language Models are Unsupervised Multitask Learners, 3 https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[6] Matt McDonnell, Benchmarking Python Content Extraction Algorithms (Jan. 29, 2015), https://moz.com/devblog/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht.

46. Upon information and belief, Defendants, when using Dragnet and Newspaper, first download and save the relevant webpage before extracting data from it. This is at least because, when they use Dragnet and Newspaper, they likely anticipate a possible future need to regenerate the dataset (*e.g.*, if the dataset becomes corrupted), and it is cheaper to save a copy than it is to recrawl all the data.

47. Because, by the time of its scraping, Dragnet and Newspaper were publicly known to remove author, title, copyright notices, and footers, and given that OpenAI employs highly skilled data scientists who would know how Dragnet and Newspaper work, Defendants intentionally and knowingly removed this copyright management information while assembling WebText.

48. A data scientist employed by Plaintiffs' counsel applied the Dragnet code to Raw Story and AlterNet URLs. As a representative sample, three results from Raw Story and three results from Alternet are attached as Exhibit 5. Of the examples Plaintiffs have examined, the resulting copies' text is in some cases completely identical to the original and in other substantively identical to the original (e.g., completely identical except for the seemingly random addition of an extra space between two words, the omission of a description or credit associated with an embedded photo, or the alteration of some formatting). Of the examples Plaintiffs have examined, every copy lacks the copyright notice information with which it was originally conveyed, while most copies lack author and title.

49. A data scientist employed by Plaintiffs' counsel also applied the Newspaper code to Raw Story and AlterNet URLs contained in OpenWebText. The data scientist applied the version of the code that enables the user not to extract author and title information based on the reasonable assumption that Defendants desired consistency with the Dragnet extractions. As a

representative sample, three results for Raw Story and three results for AlterNet are attached as Exhibit 6. Of the examples Plaintiffs have examined, the resulting copies' text is in some cases completely identical to the original and in others substantively identical to the original in the same sense as the Dragnet extractions (except that Newspaper does not randomly add spaces), while in one case a bulleted list was removed. Of the examples Plaintiffs have examined, every copy lacks the author, title, and copyright notice information with which it was conveyed to the public.

50. Plaintiffs' data scientist did not alter the article in any way other than by applying the Dragnet or Newspaper algorithm. Thus, both the removal of CMI and any trivial alterations to the article occurred simultaneously by applying the Dragnet or Newspaper algorithm to an identical copy of Plaintiffs' work.

51. The absence of author, title, and copyright notice information from the copies of Plaintiffs' articles generated by applying the Dragnet and Newspaper codes—codes OpenAI has admitted to have intentionally used when assembling WebText—further corroborates that Defendants intentionally removed author, title, and copyright notice information from Plaintiffs' copyright-protected news articles.

52. Upon information and belief, Defendants have continued to use the same or similar Dragnet and Newspaper text extraction methods when creating training sets for every version of ChatGPT since GPT-2. This is at least because Defendants have admitted to using these methods for GPT-2 and have neither publicly disclaimed their use for later version of ChatGPT nor publicly claimed to have used any other text extraction methods for those later versions.

53. Common Crawl is a data set that consists of a scrape of most of the internet created by a non-profit research institute, also called Common Crawl. ChatGPT was trained on a version of Common Crawl, in addition to the WebText and WebText2 training sets.

54. To train GPT-2, OpenAI downloaded Common Crawl data from the third party's website and filtered it to include only certain works, such as those written in English.[7] When the third party downloads the article into Common Crawl, author, title, and copyright notice information is preserved.

55. Google has published instructions on how to replicate a dataset called C4, a monthly snapshot of filtered Common Crawl data that Google used to train its own AI models. Upon information and belief, based on the similarity of Defendants' and Google's goals in training AI models, C4 is substantially similar to the filtered versions of Common Crawl used to train ChatGPT. The Allen Institute for AI, a nonprofit research institute launched by Microsoft cofounder Paul Allen, followed Google's instructions and published its recreation of C4 online.[8]

56. A data scientist employed by Plaintiffs' counsel analyzed this recreation. It contains 8,974 distinct URLs from rawstory.com. The vast majority of these URLs contain Plaintiffs' copyright-protected news articles. None of the news articles contains copyright notice information. Most lack both author and title information. In some cases, the articles' text is completely identical to the original. In other cases, the articles' text is substantively identical to the original in the same sense as the Dragnet extractions (except that the C4 algorithm does not randomly add spaces), while in still others a small number of paragraphs are omitted.

[7] Tom B. Brown et al, Language Models are Few-Shot Learners, 14 (July 22, 2020), https://arxiv.org/pdf/2005.14165.
[8] https://huggingface.co/datasets/allenai/c4.

57. Upon information and belief, both the removal of CMI and any trivial alterations to the article occurred simultaneously by applying the C4 algorithm to an identical copy of Plaintiffs' work.

58. As a representative sample, the text of three Raw Story articles as they appear in the C4 set is attached as Exhibit 7. None of these articles contains the author, title, or copyright notice information with which it was conveyed to the public.

59. In discovery, Defendants produced documents showing that Raw Story and AlterNet articles were included in Defendants' training sets. Tables listing each article's URL and the number of total occurrences across the training sets are attached as Exhibit 8 and Exhibit 9 for each article that Plaintiffs own.

60. Plaintiffs have examined a sample of the copies as they appear in the training sets. None of the copies in the examined sample contains the author, title, or copyright notice with which the original was conveyed. Apart from the removal of author, title, and copyright notice information, and the inclusion of certain symbols for formatting purposes (for example, the use of "/n/n" to denote a new paragraph), the text of the training set copies is identical to that of the originals in the sample examined by Plaintiffs.

61. Plaintiffs have not licensed or otherwise permitted Defendants to include any of their works in their training sets.

62. Defendants' actions in downloading thousands of Plaintiffs' articles without permission infringes Plaintiffs' copyright, more specifically, the right to control reproductions of copyright-protected works.

DEFENDANTS' REGURGITATION, MIMICKING, AND ABRIDGEMENT OF COPYRIGHT-PROTECTED WORKS OF JOURNALISM

63. ChatGPT offers a product to its customers that provides responses to questions or other prompts. ChatGPT's ability to provide these responses is the key value proposition of its product, one which it is able to sell to its customers for enormous sums of money, soon likely to be in the billions of dollars.

64. To train ChatGPT, Defendants retain users' chat histories with ChatGPT unless the user takes the affirmative step of disabling that feature.[9] Thus, upon information and belief, Defendants possess a repository of every regurgitation or abridgement of Plaintiffs' works apart from those whose storage users have affirmatively disabled.

65. At least some of the time, ChatGPT provides or has provided responses to users that regurgitate verbatim or nearly verbatim copyright-protected works of journalism without providing any author, title, or copyright notice information conveyed in connection with those works. Examples of such regurgitations are included in Exhibit J to the Complaint in *Daily News, LP v. Microsoft Corporation*, No. 24-cv-03285 (S.D.N.Y. Apr. 30, 2024), ECF No. 1.

66. At least some of the time, ChatGPT provides or has provided responses to users that mimic significant amounts of material from copyright-protected works of journalism without providing any author, title, or copyright notice information contained in those works. For example, if a user asks ChatGPT about a current event or the results of a work of investigative journalism, ChatGPT will provide responses that mimic copyright-protected works of journalism that covered those events, not responses that are based on any journalism efforts by Defendants.

[9] New ways to manage your data in ChatGPT (Apr. 25, 2023), https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/.

67. At least some of the time, ChatGPT memorizes and regurgitates material.[10] Defendants have publicly admitted their knowledge of this fact. Defendants have also effectively admitted that regurgitation of copyrighted works is infringement: when Plaintiffs attempted to obtain the same regurgitations set forth in the *Daily News* case using the same methodology, Plaintiffs received in one instance a message stating, "I'm sorry, but I can't generate the original ending for the article or any copyrighted content." Thus, upon information and belief, Defendants have recently changed ChatGPT to reduce regurgitations for copyright reasons.

68. At least some of the time, ChatGPT provides or has provided responses to users that abridge copyright-protected works of journalism without providing any author, title, or copyright notice information conveyed in connection with those works. Examples of such abridgements are included in Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc.*, No. 24-cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. For instance, in the fourth example, the ChatGPT abridgement reproduces, verbatim, nine consecutive paragraphs of text (minus one sentence) from the original article, which can be found at https://www.motherjones.com/politics/2024/01/100-bill-crime-corruption-treasury-tax-evasion/.

69. When earlier versions of ChatGPT (up to and including ChatGPT 3.5-turbo) abridge a copyright-protected news article in response to a user prompt, they draw from their training on the article. During training, the patterns of all content, including copyright-protected news articles, are imprinted onto the model. That imprint allows the model to abridge the article.

70. When later versions of ChatGPT abridge a copyright-protected news article in response to a user prompt, they find the previously downloaded article inside a database called a

[10] OpenAI and journalism (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/.

search index using a method called synthetic searching or retrieval-augmented generation ("RAG"). Upon information and belief, they make another copy of the article in the memory of their computing system and use their LLM or other programming to generate an abridgement by applying the model or other programming to the text of the article.

71. Plaintiffs' articles are not merely collections of facts. Rather, they reflect the originality of their authors in selecting, arranging, and presenting facts to tell compelling stories. They also reflect the authors' analysis and interpretation of events, structuring of materials, marshaling of facts, and the emphasis given to certain aspects.

72. An ordinary observer of a ChatGPT abridgement of copyright-protected news articles would conclude that the abridgements were derived from the articles being abridged.

73. In response to prompts seeking an abridgement of an article, ChatGPT will typically provide a general abridgement of such an article, on the order of a few paragraphs. In some instances, the initial response will summarize the article in substantial detail. Further, when prompted by the user to provide more information about one or more aspects of that abridgement, ChatGPT will provide additional details, often in the format of a bulleted list of main points. If prompted by the user to provide more information on one of more of those points, ChatGPT will provide additional details. In some instances, however, ChatGPT will provide a bulleted list of main points in response to an initial prompt seeking an abridgement.

74. A ChatGPT user is capable of obtaining a substantial abridgement of a copyright protected news article through such series of prompts, and in some instances, further prompts designed to elicit further summary are even suggested by ChatGPT itself. *See* Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc.*, No. 24-cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. These abridgements lack copyright notice

information conveyed in connection with the work, and sometimes lack author information. They sometimes link to webpages that do not belong to the news organization that owns the article and that do not contain the news organization's copyright management information.

75. Thus, upon information and belief, abridgements from earlier versions of ChatGPT lack copyright notice and typically author information because Defendants intentionally removed that information from the ChatGPT training sets.

76. Further, the abridgements from later versions of ChatGPT lack copyright notice and typically author information. Upon information and belief, this is because Defendants intentionally removed them either when initially storing them in computer memory or when generating results by employing RAG.

77. On April 22, 2024, Plaintiffs served a Request for Production on Defendants seeking all chat responses that included any portion of any articles published on rawstory.com or alternet.org. That same day, Plaintiffs also served a Request for Production on Defendants seeking all chat responses that regurgitated any articles from rawstory.com or alternet.org.

78. Defendants agreed to produce this information many months ago, but despite Plaintiffs' diligent follow-ups, they have not yet done so. Analysis of this information would show the extent to which Defendants provided regurgitations or other outputs that sufficiently disseminated Plaintiffs' works or abridgement of those works.

DEFENDANTS' INTENTIONAL REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION FROM PLAINTIFFS' WORKS IN THEIR TRAINING SETS

79. ChatGPT does not have any independent knowledge of the information provided in its responses. Rather, to service Defendants' paying customers, ChatGPT instead repackages, among other material, the copyrighted journalism work product developed by Plaintiffs and others at their expense.

80. When providing responses, ChatGPT gives the impression that it is an all-knowing, "intelligent" source of the information being provided, when in reality, the responses are frequently based on copyrighted works of journalism that ChatGPT simply mimics.

81. If ChatGPT was trained on works of journalism that included the original author, title, and copyright information, ChatGPT would have learned to communicate that information when providing responses to users unless Defendants trained it otherwise.

82. Based on the publicly available information described above, thousands of Plaintiffs' copyrighted works were included in Defendants' training sets without the author, title, and copyright notice information that Plaintiffs conveyed in publishing them.

83. Based on the publicly available information described above, including Defendants' admission to using the Dragnet and Newspaper extraction methods, which remove author, title, and copyright notice information from copyright-protected news articles published online, Defendants intentionally removed author, title, and copyright notice information from Plaintiffs' copyrighted works in creating ChatGPT training sets.

DEFENDANTS' ACTUAL AND CONSTRUCTIVE KNOWLEDGE OF THEIR VIOLATIONS

84. Defendants have acknowledged that use of copyright-protected works to train ChatGPT requires a license to that content. and, in some instances. Recognizing that obligation, Defendants have entered into agreements with large copyright owners such as Associated Press, the Atlantic, Axel Springer, Dotdash Meredith, Financial Times, News Corp, and Vox Media to obtain licenses to include those entities' copyright-protected works in Defendants' LLM training data.

85. Defendants are also in licensing talks with other copyright owners in the news industry, but have offered no compensation to Plaintiffs.

86. In a May 29, 2024 interview, OpenAI's Chief of Intellectual Property and Content, Tom Rubin, stated that these deals focus on "the display of news content and use of the tools and tech," and are thus "largely not" about training.[11] This admission, while qualified, confirms that these deals involve training, at least in part.

87. Defendants created tools in late 2023 to allow copyright owners to block their work from being incorporated into training sets. This further corroborates that Defendants had reason to know that use of copyrighted material in their training sets without permission or license is copyright infringement.

88. The creation of such tools also corroborates that Defendants had reason to know that their copyright infringement is enabled, facilitated, and concealed by their removal of author, title, and copyright notice information from their training sets.

89. Defendants had reasonable grounds to know that the removal of author, title, and copyright notice information from copyright-protected works and their use in training ChatGPT would result in ChatGPT providing responses to ChatGPT users that incorporated or regurgitated material verbatim from copyrighted works in creating responses to users, without revealing that those works were subject to Plaintiffs' copyrights. This is at least because Defendants were aware that ChatGPT responses are the product of its training sets and that ChatGPT would not know any author, title, and copyright information that was not included in training sets.

90. Upon information and belief, Defendants had reason to know that the removal of author, title, and copyright notice information from copyright-protected works used in synthetic searching would result in ChatGPT providing responses to ChatGPT users that abridged or

[11] Charlotte Tobitt, OpenAI content boss: 'Incumbent' on us to help small publishers, not just the giants, *PressGazette* (May 30, 2024), https://pressgazette.co.uk/platforms/openai-tom-rubin-publishers-news/.

regurgitated material verbatim from copyrighted works in creating responses to users, without revealing that those works were subject to Plaintiffs' copyrights. This is at least because Defendants were aware that later versions of ChatGPT's responses to prompts are the product of the articles encoded in their computer memory, from which, upon information and belief, Defendants removed author, title, and copyright notice information.

91. Defendants had reason to know that users of ChatGPT would further distribute the results of ChatGPT responses. This is at least because Defendants promote ChatGPT as a tool that can be used by a user to generate content for a further audience.

92. Defendants had reason to know that users of ChatGPT would be less likely to distribute ChatGPT responses if they were made aware of the author, title, and copyright notice information applicable to the material used to generate those responses. This is at least because Defendants were aware that at least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement.

93. Defendants had reason to know that ChatGPT would be less popular and would generate less revenue if users believed that ChatGPT responses violated third-party copyrights or if users were otherwise concerned about further distributing ChatGPT responses. This is at least because Defendants were aware that they derive revenue from user subscriptions, that at least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement, and that such users would not pay to use a product that might result in copyright liability or did not respect the copyrights of others.

94. If a commercial user of ChatGPT is sued for copyright infringement, Defendants have committed to paying the user's costs in defending against the infringement claim, and to indemnifying the user for an adverse judgment or settlement. These commitments apply only if

the user uses the product as advertised. In particular, OpenAI's "Copyright Shield" does not apply if the user "disabled, ignored, or did not use any relevant citation, filtering or safety features or restrictions provided by OpenAI."[12] Thus, Defendants know or have reason to know that ChatGPT users are capable of infringing and likely to infringe copyright even when used according to terms specified by Defendants.

95. Defendants intend in part for ChatGPT to replicate how ordinary English speakers express themselves. When ordinary English speakers are not conveying copyright-protected works, they do not include copyright management information—especially copyright notices. Had ChatGPT been trained on Plaintiffs' and others' copyright-protected works that include this copyright management information, it would have falsely learned that ordinary English speakers convey copyright management information in situations when they do not. To avoid this result, Defendants had a choice between removing the copyright management information at the outset or retraining ChatGPT not to emit the copyright management information after it had incorrectly learned how English speakers normally express themselves. Upon information and belief, Defendants chose to remove the copyright management information at the outset, at least because doing so involves fewer computational resources and therefore is far less expensive than retraining. Thus, because Defendants infringed Plaintiffs' copyright by using Plaintiffs' articles to train ChatGPT, Defendants removed Plaintiffs' copyright management information from its copyright-protected articles knowing, or having reasonable grounds to know, that doing so would facilitate their own training-based infringing conduct.

[12] https://openai.com/policies/service-terms/.

96. Defendants' unauthorized copying of Plaintiffs' works into Defendants' training sets and search indices is facilitated by the removal of author, title, and copyright notice information because copying less data requires fewer computational and storage resources.

DEFENDANTS' CONTINUING VIOLATIONS

97. Upon information and belief, Defendants have continued to unlawfully copy, regurgitate, abridge, and remove author, title, and copyright notice information from Plaintiffs' copyright-protected works up to the present date, or at least until Plaintiffs implemented the exclusion protocols on October 27, 2023, that Defendants released in August 2023 allowing websites to opt out of OpenAI's web crawling.

98. ChatGPT has emitted significant material from copyright-protected works of journalism that significantly postdate the WebText and WebText2 training sets. Examples are contained in Exhibit 11 to the First Amended Complaint in *The Center for Investigative Reporting, Inc. v. OpenAI, Inc.*, No. 24-cv-4872 (S.D.N.Y. Sept. 9, 2024), ECF No. 88-14. ChatGPT could not have produced this material without Defendants' copying the original articles and storing them in computer memory, including in training sets created by ChatGPT 3.5-turbo and earlier, and search indices for RAG purposes.

99. In addition, each successive GPT model has had orders of magnitude more parameters than the last. For instance, GPT-4 is reported to have 1.8 trillion parameters,[13] a tenfold increase from the 175 billion parameters used to train GPT-3.[14] Because adding more parameters requires training on more data, it is unlikely that Defendants would have foregone including

[13] Maximilian Schreiner, GPT-4 architecture, datasets, costs and more leaked, The Decoder (July 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.
[14] Tom B. Brown et al, Language Models are Few-Shot Learners, 5 (July 22, 2020), https://arxiv.org/pdf/2005.14165.

Plaintiffs' articles in their more recent training sets. Thus, upon information and belief, Defendants continue to include Plaintiffs' articles in their training sets up to the present date.

100. Further, Defendants' adoption of a tool in August 2023 to allow website owners to block web crawling would have been unnecessary they were not continuing to copy content from the internet, including Plaintiffs' copyright-protected works, as they had done in the past.

101. According to OpenAI's Chief of Intellectual Property and Content, each of OpenAI's models is "trained from scratch."[15] Thus, when assembling new training sets, OpenAI recrawls the same articles it included in past training sets.

102. As alleged above, upon information and belief, Defendants have continued to use the same or similar Dragnet and Newspaper text extraction methods when creating training sets for every version of ChatGPT since GPT-2. Thus, upon information and belief, they have continued to remove author, title, and copyright notice information from Plaintiffs' copyright-protected articles up to the present, including but not limited to Plaintiffs' articles that are contained in Defendants' training sets created in the past three years.

Count I – Violation of 17 U.S.C. § 1202(b)(1)

103. The above paragraphs are incorporated by reference into this Count.

104. Plaintiffs are the owners of copyrighted works of journalism that contain author, title, and copyright notice information.

105. Defendants created copies of Plaintiffs' works of journalism with author information removed and included them in training sets used to train ChatGPT.

[15] Charlotte Tobitt, OpenAI content boss: 'Incumbent' on us to help small publishers, not just the giants, PressGazette (May 30, 2024), https://pressgazette.co.uk/platforms/openai-tom-rubin-publishers-news/.

106. Defendants created copies of Plaintiffs' works of journalism with title information removed and included them in training sets used to train ChatGPT.

107. Defendants created copies of Plaintiffs' works of journalism with copyright notice information removed and included them in training sets used to train ChatGPT.

108. Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright notice information would induce ChatGPT to provide responses to users that incorporated material from Plaintiffs' copyright-protected works, and abridged or regurgitated copyright-protected works verbatim or nearly verbatim.

109. Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright notice information would induce ChatGPT users to distribute or publish ChatGPT responses that utilized Plaintiffs' copyright-protected works of journalism that such users would not have distributed or published if they were aware of the author, title, and copyright notice information.

110. Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright notice information would enable copyright infringement by Defendants, ChatGPT and ChatGPT users.

111. Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright notice information would facilitate copyright infringement by Defendants, ChatGPT and ChatGPT users.

112. Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright notice information would conceal copyright infringement by Defendants, ChatGPT, and ChatGPT users.

PRAYER FOR RELIEF

Plaintiffs seek the following relief:

(i) Either statutory damages or the total of Plaintiffs' damages and Defendants' profits, to be elected by Plaintiffs;

(ii) An injunction requiring Defendants to remove all copies of Plaintiffs' copyrighted works from which author, title, or copyright notice information was removed from their training sets and any other repositories;

(iii) An injunction prohibiting the unlawful conduct alleged above;

(iv) An injunction ordering the destruction of all GPT or other LLMs and training sets that incorporate Plaintiffs' works from which author, title, or copyright notice has been removed; and

(v) Attorney fees and costs.

JURY DEMAND

Plaintiffs demand a jury trial.

RESPECTFULLY SUBMITTED,

/s/ Stephen Stich Match

Jon Loevy (*pro hac vice*)
Michael Kanovitz (*pro hac vice*)
Lauren Carbajal (*pro hac vice*)
Stephen Stich Match (No. 5567854)
Matthew Topic (*pro hac vice*)
Thomas Kayes (*pro hac vice*)
Steven Art (*pro hac vice*)
Kyle Wallenberg (*pro hac vice*)

LOEVY & LOEVY
311 North Aberdeen, 3rd Floor
Chicago, IL 60607
312-243-5900
jon@loevy.com
mike@loevy.com
carbajal@loevy.com
match@loevy.com
matt@loevy.com
kayes@loevy.com
steve@loevy.com
wallenberg@loevy.com

November 21, 2024